Intelligent Data Analysis 19th October, 2016

Homework 1.2

Professor: Arminda Moreno Díaz

Students: Jose Luis Contreras Santos and Antonio Javier González Ferrer

Introduction

This document aims to extend on the analysis of the *diamonds* dataset from ggplot2, which was started in Homework 1.1 of this series of exercises, performing a basic exploration of the dataset, obtaining detailed descriptive statistics about some of the variables it contains.

Exercise 1.2

Bivariate analysis: Price and Color. The first step is to analyze how price is distributed across the different colors. For this purpose, the following boxplot has been generated, using a logaritmic scale on the y-axis for the sake of simplicity in the representation.

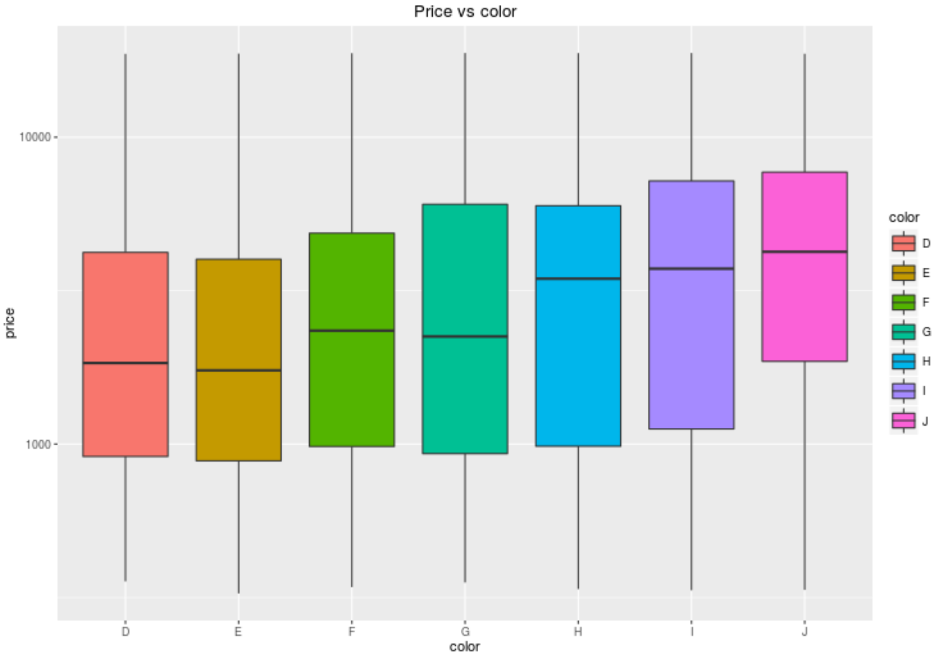


Figure 1: Boxplot representing the relationship between *price* and *color*.

The plot illustrates perfectly what the description of the dataset states: colors are ordered from worst to best, with J being the most expensive color and D the cheapest. Although it is not strictly true for all colors (diamonds of color E, for example, have a slightly lower mean price than those of color D, which is supposedly worse), the trend is clear: diamonds of a *better* color have a greater chance of having high price tags than those of less attractive colors.

Univariate analysis: Price. The univariate quantitative variable to analize is the '*price*' of the diamonds. First of all, let us plot the histogram of the variable in order to observe graphically its distribution:

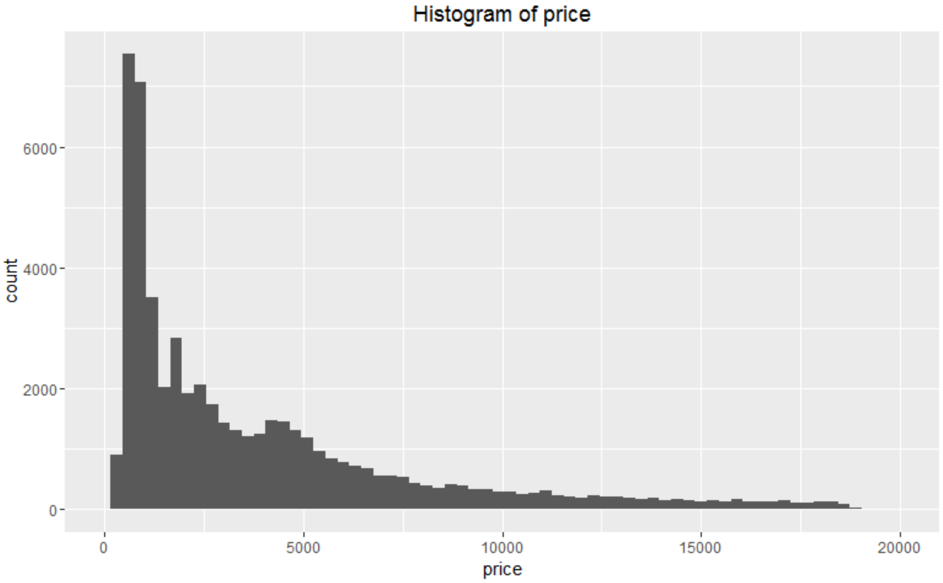


Figure 2: Histogram of the univariate variable '*price*'. At first glance, we guess that data is not centered and has a large variance, a left-skewed distribution and a leptokurtic kurtosis.

The label of Figure 2 speculates about different measures of the '*price*' distribution. Let us check our assumptions analitically:

- **Center**: The mean is the most common measure of center. In this case $\mu = 3932.8$. However, the mean is affected by extreme values so it may not be the best measure of center to use in a skewed distribution. Instead, the **median** would probably represent the best measure of center: $M = 2401$.

- **Dispersion**: The data values are fairly dispersed around the mean, being the variance of the variable $\sigma^2 = 15915629$. To make this value more interpretable, let us rather consider the **standard deviation** $\sigma = 3989.44$, since it is expressed in the same units as the median.

- **Skewness**: The skewness is related to the assymetry of the distribution, being in this case $g_1 = 1.61835$ which represents that the mass of the distribution is concentrated on the left side of the figure.

- **Kurtosis**: The kurtosis measures the "tailedness" of the distribution, that is to say, it measures the variance of the outliers with respect to the mean. Since kurtosis $= 5.177383$, the distribution is heavy-tailed and therefore leptokurtic.

Based on the skewness and kurtosis, we are going to use the Jarque-Bera test to check normality, which uses as H_0 if the distribution is normal. We obtain a $\chi^2 = 34201$ and a p-value $< 2.2e - 16$, hence there exists strong evidence to reject H_0. A normal model is not plausible for the distribution of '*price*'. However, a data transformation in the variable could allow to use a normal mode under its distribution, although transforming the variable may lead to a loss in interpretability. Thanks to the suggestions given in the lectures, where we saw several initial distributions and their transformations to improve normality, we think that a logarithmic transformation would enhance normality[1]. Aplying again the Jarque-Bera test, we get $\chi^2 = 2823.1$ and a p-value $< 2.2e-16$. Although there is still strong evidence of the distribution not being normal, the χ^2 coefficient has been dramatically reduced. Figure 3 plots the empirical quantiles of the variable *'price'* before and after applying the log transformation compared to the normal distribution.

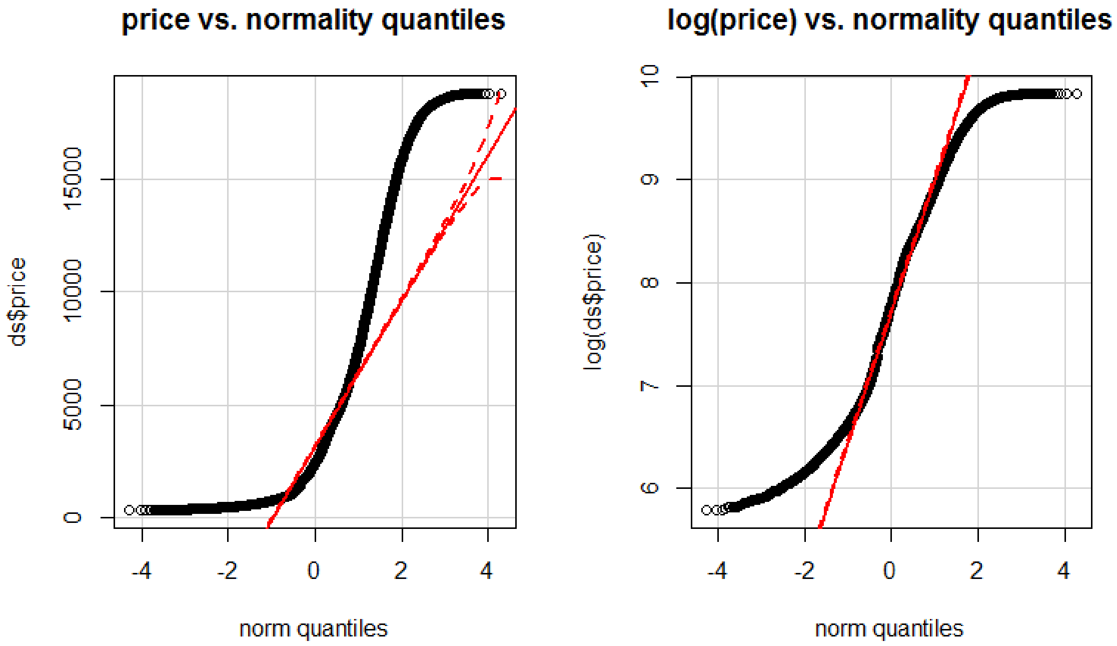
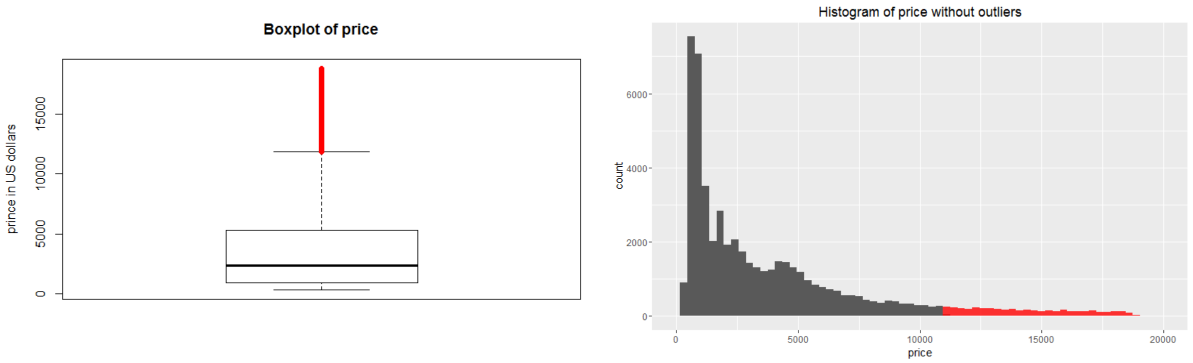


Figure 3: a) *qqPlot* before transformation. b) *qqPlot* after apply log transformation.

Normality is a favorable assumption when we apply models (e.g. regression analysis) to the sample in order to predict the behaviour of the population. The lack of this normality can occur due to the existence of outliers in the data. Sometimes, it is interesting to study the tradeoff between the loss of information when the outliers are removed and the gain in normality of the distribution.

The boxplot of Figure 4(a) shows the distribution of the '*price*' variable and its outliers. In total, there are 3538 outliers (which account for a 6% of the data), where the prices range from 11000 up to 19000. A comparison between the initial distribution and the distribution after the deletion of the outliers can be seen in Figure 4(b). Nevertheless, since the removal of outliers still does not ensure normality and the χ^2 coefficient only decreases to 1535.4 (compared to 2823.1), we decided to maintain the outliers, thus avoiding the loss of information.

[1] After we took this decision, we used the maximum likelihood-like approach of Box and Cox to check if the transformation could be improved. Since the *powerTransform* function selected a transformation of $\lambda = -0.067$, we decided to maintain the log transformation for interpretability reasons.



(a) Boxplot distribution of ‘*price*‘. Outliers are marked in red.

(b) Distribution of ‘*price*‘ after the ourliers were eliminated (in red).

Figure 4: Univariate outliers analysis of the variable ‘*price*‘.

Bivariate quantitative analysis: Carat and Price. Due to memory limitations, we have selected a subset of the original data to analyze the normality of the joint bivariate distribution between the carat and the price. Sometimes, diamond cutters are interested in studying the evolution of gems within a certain range of carat. We have chosen diamonds between 0.95 and 0.97, which correspond to a collection of 227 observations. Figure 5 a) shows the *qqPlot* of the distribution after applying the Mardia's Multivariate Normality test. Either in this case, the joint bivariate distribution shows no evidence of being normal, due to a high kurtosis value (kurtosis $= -1.098$ and p-value $= 0.2719442$). Nonetheless, after applying the maximum likelihood-like transformation of Box and Cox, there does not exist evidence against the null hyphotesis (p-value < 0.185447) and hence the distribution follows normality.

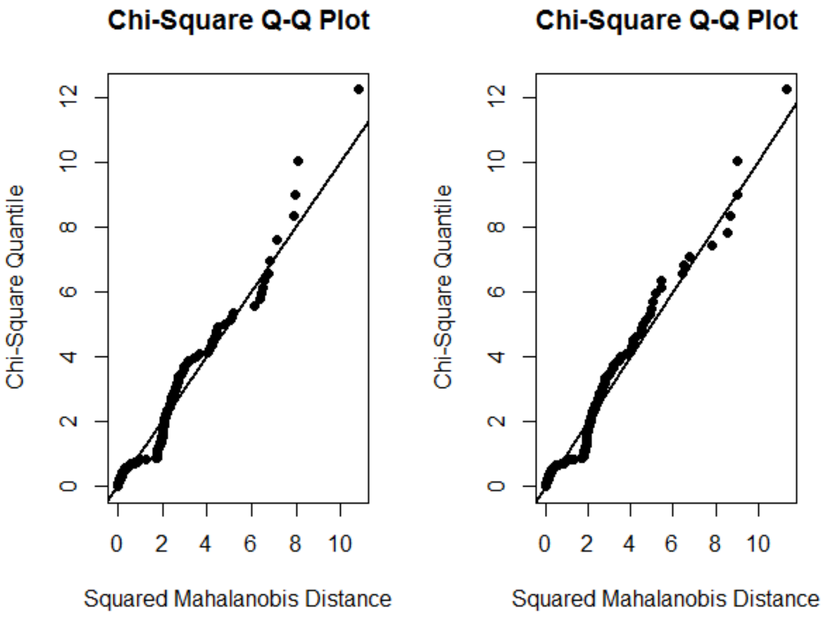
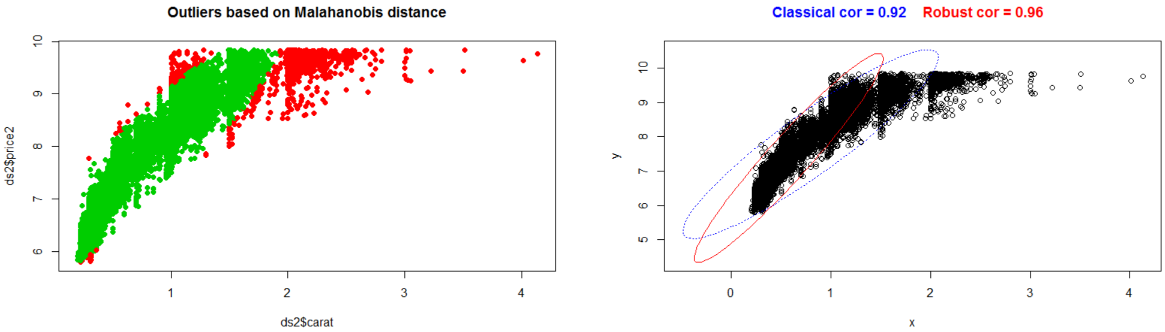


Figure 5: Normality study of diamond's carat between 0.95 and 0.97. a) *qqPlot* before transformation. b) *qqPlot* after Box Cox transformation ($\lambda_1 = -2.714051, \lambda_2 = 6.429170$).

To conclude, we have analyzed the outliers of the original joint bivariant distribution from two differents points of view. Firstly, we have computed the outliers based on the Mahalanobis distance (Figure 6 a)). The Mahalanobis distance is a multidimensional generalization of the idea of measuring how many standard deviations away a point is from the mean of the distribution, taking into account the correlation of the variables. The algorithm has detected 1394 points as candidate outliers. Secondly, we have compared in Figure 6 b) this robust method against the classical bivariate correlation. Notice the robust correlation is more natural than the classical correlation, according to the distribution.

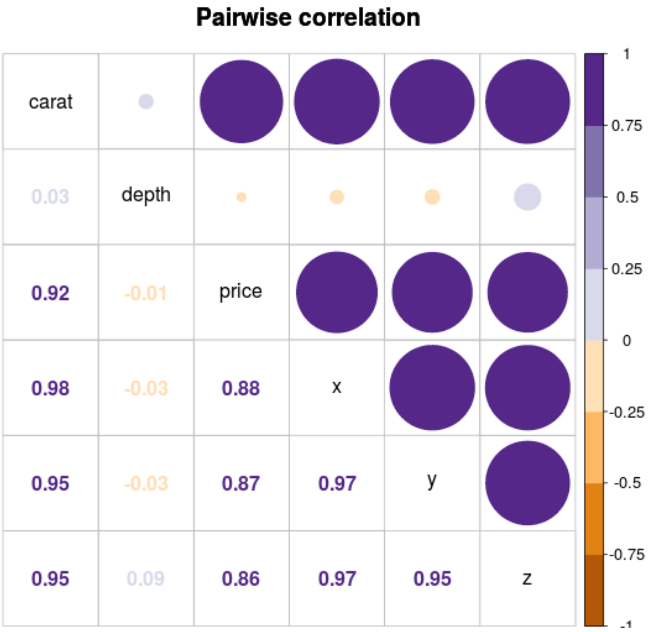


(a) Outliers detection based on Mahalanobis distribution. The outliers are the red points.

(b) Comparison between classical correlation (blue) and robust correlation (red) outliers detection.

Figure 6: Outliers of the joint bivariate distribution between '*carat*' and '*price*'.

Study of linear relationships. The last step to finish our analysis of the *diamonds* dataset is to study the linear relationships between its quantitative variables. The variables to be inspected are *'price'*, *'carat'* (diamond weight), *'x'*, *'y'* and *'z'* dimensions, and *'depth'*.



carat					
0.03	depth				
0.92	-0.01	price			
0.98	-0.03	0.88	x		
0.95	-0.03	0.87	0.97	y	
0.95	0.09	0.86	0.97	0.95	z

Figure 7: Pairwise correlations between the aforementioned variables.

The most prominent feature seen in the pairwise correlation plot (Figure 7) is the inexistent linear relation between ‘*depth*‘ and the rest of the variables. This, however, does not come as a surprise, as we already know this number is obtained via a non-linear combination of ‘*x*‘, ‘*y*‘ and ‘*z*‘[2]. Results also show a strong linear dependence between a diamond's weight, price, and dimensions. Yet, there are reasons to suspect that some of these results may be biased due to intrinsic relations within the data. Specially, the link between price and each of the dimensions is most probably affected by the influence the latter have on the gem's weight. Luckily for us, this can be analyzed by using the matrix of partial correlations, which we will generate before reaching premature conclusions.

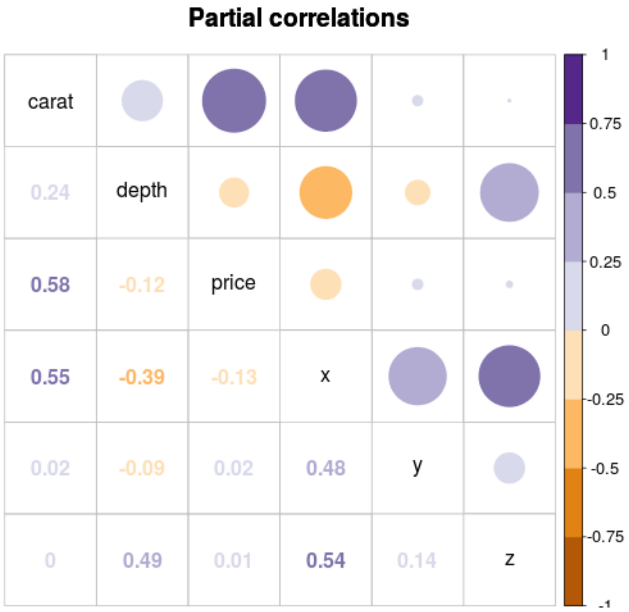


Figure 8: Matrix of pairwise correlations.

Partial correlations (Figure 8) show weak linear dependence between the variables. Weight and price are indeed correlated, but not as much as the previous results suggested (color, as analyzed before, and probably also cut quality are influencing price too), and the majority of the high correlation scores are now much lower. As a side note, it is interesting to see how ‘*x*‘ and the ‘*carat*‘ parameter are related, but the rest of dimensions are not. The ‘*x*‘ is also somehow related to the other dimensions, which is most likely due to most diamonds having a relatively similar shape.

The coefficients of determination can also give us a good insight into linear relationships within the data. As the Table 1 below displays, important linear relations are present in this set. Specifically, and as already studied before, ‘*carat*‘ and the three dimensional variables are the most linearly explained by the other variables. Slightly less influenced, but still scoring high, is the *price* tag.

[2] Total depth percentage, calculated as

$$D = \frac{z}{mean(x, y)} = \frac{2z}{x + y}$$

	carat	depth	price	x	y	z
R-squared	0.969	0.296	0.856	0.982	0.951	0.957

Table 1: Coefficient of determination for each of the analyzed variables.

Until this point, our analysis shows strong linear relations in this dataset. Further proof of this fact can be found in the effective dependence coefficient of the R matrix: $D(R) = 0.8445259$. This means that, altogether, linear dependences explain 84% of the variability of the data.

To complete the study, the eigenanalysis of the R matrix has been performed. The lowest of its eigenvalues (Table 2) is close enough to zero to justify checking the corresponding eigenvector in order to find linear relationships, which is as follows: $[-0.4241, 0.0549, 0.0818, 0.8293, -0.2095, -0.2804]$. Thus, we can see that variables 1, 5 and 6 (that is, *'carat'*, *'y'* and *'z'*) hold linear relations with variable 4 (x).

Eigenvalues	4.7265	1.0109	0.1762	0.0404	0.0334	0.0124

Table 2: Eigenvalues of the R matrix